2102531 System Identification

Time Series Models of Stock Price
Semester 1/2018

Maytus Piriyajitakonkij
Jitkomut Songsiri
Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University

November 25, 2018

Abstract

Stock market movements are difficult to predict precisely. Prediction of price in the stock markets with high accuracy make a tremendous advantage for investors. There are two types of stock analysis. First is fundamental analysis which evaluate the price using the most fundamental financial level of company. Second is technical analysis which evaluate the price by gathering previous trading activity in markets, such as opening price, closing price, high price, low price and volume, and creating predictive models. In this paper we use many popular predictive models, i.e. ARIMA, ARIMAX and Artificial Neural Network, to predict stock price. We found that neural networks is the most powerful tool in this paper for stock price prediction.

1 Introduction

Stock market have many variables (opening price, closing price, high price, low price and volume). Therefore, we can choose the variables in many different ways and we can extract the useful information from data using many different models. The description of stock data is in the subsection below.

1. Opening price is the first price at which a security is traded at the opening time of the market.
2. Closing price is the last price at which a security is traded at the closing time of the market.
3. High price is the maximum price at which a security is traded throughout the day.
4. Low price is the minimum price at which a security is traded throughout the day.
5. Volume is the total trading quantity (buying and selling) at which a security is traded throughout the day.

In many research [1, 2, 3], they use closing price to represent the stock market. The models which are used for prediction are various type from time series models which is ARIMA to Deep learning, i.e. Artificial Neural Networks and Recurrent Neural Networks. Therefore, we need to evaluate each model to find out what is suitable model for predict stock price. In this paper, we also choose closing price to represent stock price in the market because it is a reasonable representation for all trading activity in a day. We focus on short term prediction which use the price from previous days to predict the price in the next day.

2 Background

We use two types of model in this literature to predict stock price.

1. Time Series Models are statistical methods which are used for analyzing data in time point and finding profitable information in historical data that can predict some values in the future. Model formulation for time series data depend on type of stochastic process in data. There are some widely used models for stock price prediction.

   (a) Time series model for stationary process of which pattern in statistical properties can be found, such as constant mean.

   - Autoregressive Moving Average Model (ARMA) provide a description of stationary process. There are two polynomial terms. First is regressive term which contains previous time data, and second is moving average term which contains previous and present time errors.

   $$A(L)y(t) = C(L)e(t) \tag{1}$$

   - Autoregressive Moving Average with Exogenous Inputs (ARMAX) is similar to ARMA Model but it has an extra term which contains inputs at various times from previous.

   $$A(L)y(t) = B(L)u(t) + C(L)e(t) \tag{2}$$

   (b) Time series model for non-stationary process of which pattern in statistical properties can't be found directly.

   - Autoregressive Integrated Moving Average Model (ARIMA) provide a description of stationary process. There are two polynomial terms. First is regressive term which contains previous time data, and second is moving average terms which contain previous and present time errors. We choose $y(t)$ as daily closing prices in this model.

   $$A(L)(1-L)^d y(t) = C(L)e(t) \tag{3}$$

   - Autoregressive Integrated Moving Average with Exogenous Inputs (ARIMAX) is similar to ARMA Model but it has a extra term which contains inputs at various times from previous.

   $$A(L)(1-L)^d y(t) = B(L)u(t) + C(L)e(t) \tag{4}$$

   where

   $$A(L) = I - (A_1 L + ... + A_p L^p)$$
   $$B(L) = B_1 L + ... + B_n L^n$$
   $$C(L) = I + C_1 L + ... + C_q L^q$$

2. Deep Learning is subsets of machine learning model. They are inspired by information processing and transmission in nervous systems, such as human brain. Their applications are widely in computer vision, natural language processing and automatic speech recognition. Moreover, many recent research [1, 2] use deep learning to predict stock price.

   - Artificial Neural Network (ANN) is computational model which is inspired from the way human brain functions and structures. This model has an advantage in dealing with nonlinear system. It can recognize complex thing, such as human face, therefore we will apply this model to predict the stock price.

   $$\hat{y} = f(X) \tag{5}$$

   where
   $\hat{y}$ is predicted output from neural networks.
   $X \in \mathbb{R}^n$ is the input vector of neural networks.
   $f(X)$ is the nonlinear function which represents the neural networks.

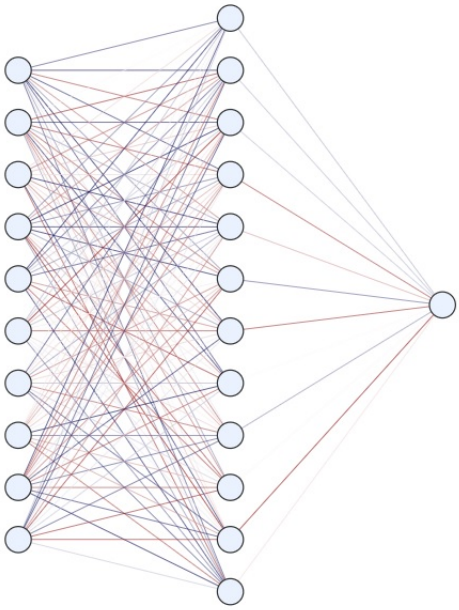

Figure 1: Fully connected ANN with 10 input nodes and 1 output node. Created by: alexlenail.me

Model Selection Criteria

There are many criteria to evaluate ARIMA and ANN for determining which is the best. According to [1], we use RMSE, AIC and BIC on training set for model evaluation and selection for ARIMA and only RMSE for ANN. According to [4].

- Root-mean-square error (RMSE) is used to compare the deviation of predicted values from observed values, and can be representative the model accuracy.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{6}$$

- Akaike information criterion (AIC) is used to determine the best model. This criterion can allow us to trade off betweern the number of parameters and mean-square error.

$$AIC = -2\mathcal{L}_k + 2k \tag{7}$$

- Bayesian information criterion (BIC) is similar to AIC in sense of its purpose. It perform well at getting the right order in large samples, while AIC are good in smaller samples.

$$BIC = -2\mathcal{L}_k + 2k\log n \tag{8}$$

where

n is the number of samples.

k is the number of regression parameters in the model.

$\mathcal{L}_k$ the maximized value of the likelihood function.

3 Problem statement

1. The stock data used in this paper are KBANK historical daily stock prices obtained from 1 January 2008 to 26 October 2018. There are 2626 observations
2. The price used to represent the market is closing price.
3. The prediction is daily (time step is one day) which use the data from previous days which are not only closing price in some models, i.e. $y(t), y(t-1), ...,$ to predict the closing price on the next day ($\hat{y}(t+1)$).
4. The models used for stock prediction are ARIMA, ARIMAX and three-layer Neural Networks with the different inputs. ARIMA and first ANN uses only daily closing price, ARIMAX and second ANN will be added more inputs which we will select them in methodology section.
5. The criterion for evaluating performance of the models to is root-mean-square-error (RMSE) on test set.

4 Methodology

The methods which are used in this paper to develop predictive model for stock price prediction are explained in the subsection below. The tools and libraries used in this paper are Jupyter Notebook for editing python code, Statsmodels for implementing time series models and Scikit Learn for implementing artificial neural network. According to above explaination, the closing price is representative of KBANK stock price. There are four models we use for prediction.

1. **ARIMA(p, d, q)**

   We choose 161 observations from 1 January 2018 to 28 August 2018 to be training set and 41 observations from 29 August 2018 to 26 October 2018 to be test set in ARIMA because the data which is obtained from a distant past is not capability to indicate present and this model has a few parameters to train. Figure 2 show the overview of stock price. It looks like a random walk which is nonstationary process. We find the ACF and PACF. Figure 3 show that the ACF decay slowly which is the characteristic of nonstationary process. We try to convert nonstationary process to stationary process by differencing. First difference stock price is shown in Figure 4 with ACF and PACF in Figure 5. Therefore, it looks close to white noise which is wide-sense stationary process that we can handle it with time series models.

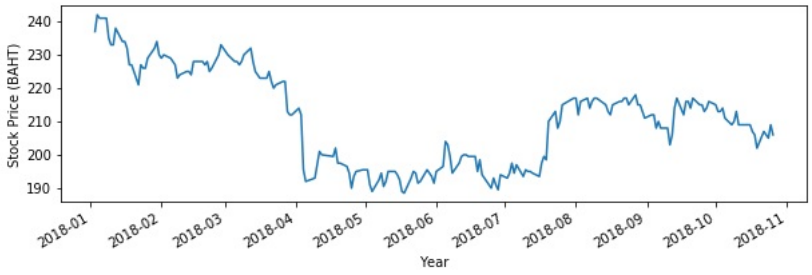

Figure 2: KBANK closing price index

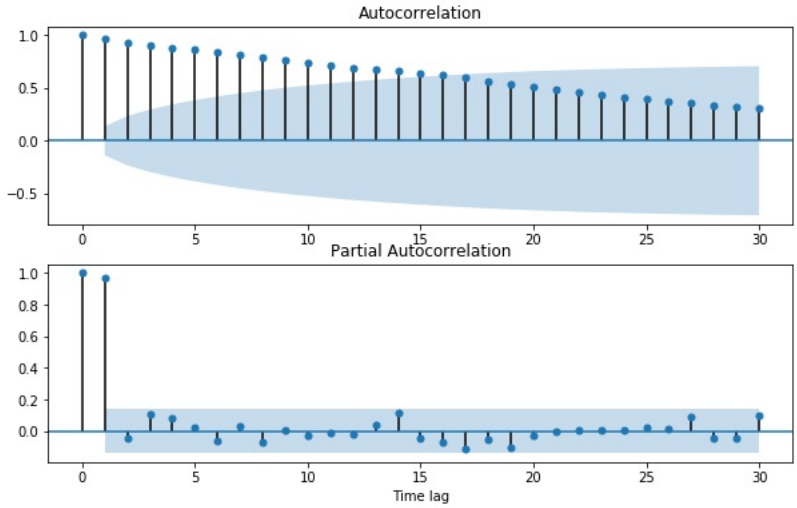

Figure 3: The correlogram of KBANK closing price index

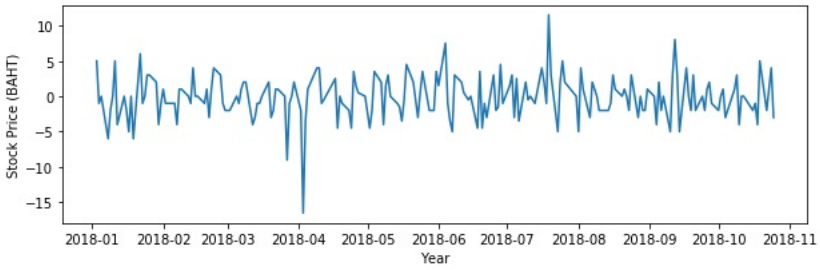

Figure 4: KBANK closing price index after first differencing

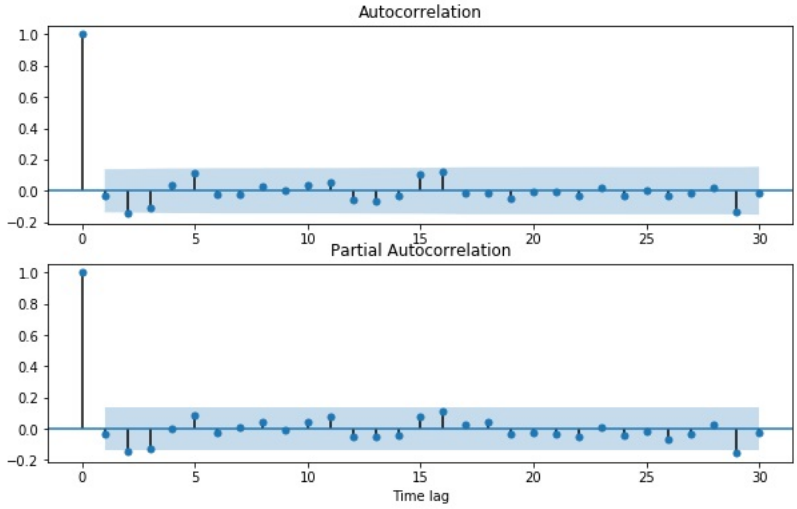

Figure 5: The correlogram of KBANK closing price index after first differencing

ARIMA	Training AIC	Training BIC	Training RMSE
(0, 1, 1)	827	836	3.15
(0, 1, 2)	825	837	3.11
(1, 1, 0)	827	836	3.15
(1, 1, 1)	828	840	3.14
(1, 1, 2)	826	842	3.10
(2, 1, 0)	825	837	3.11
(2, 1, 1)	826	841	3.09
(3, 1, 0)	**825**	**840**	**3.09**
(3, 1, 1)	827	845	3.09

Table 1: Statistical results of ARIMA models for KBANK stock price prediction

We train many different models using grid search which is shown in Table 1. The parameters (p,d,q) represent autoregressive terms (p), moving average terms(q) and d^{th} time differencing. According to [4], range of p is selected from PACF closing price after first differencing which have value on lag = 3, therefore the maximum order of autoregressive term is 3. Later, range of q is selected from ACF closing price after first differencing which have value on lag = 2, therefore the maximum order of moving average term is 2. Finally, d is one because the closing price after first differencing seem white noise which is wide-sense stationary process.

We found that ARIMA(3, 1, 0) had the smallest RMSE and AIC on training set. It is accpeptable that only BIC is not the smallest, but it is not much large. We choose it to be the candidate for model comparison.

2. **ARIMAX(p, d, q)**

We choose inputs by cross correlation analysis. First, we apply differencing on Volume, Opening price, High price and Low price. Then, we find the cross correlation between the these variables and closing price after first differncing. The results are demonstrated in Figure 6. Volume after differencing on the past is uncorrelated with closing price. In contrast, opening, high and low price cross correlation are slightly correlated with closing price on time lag 2.

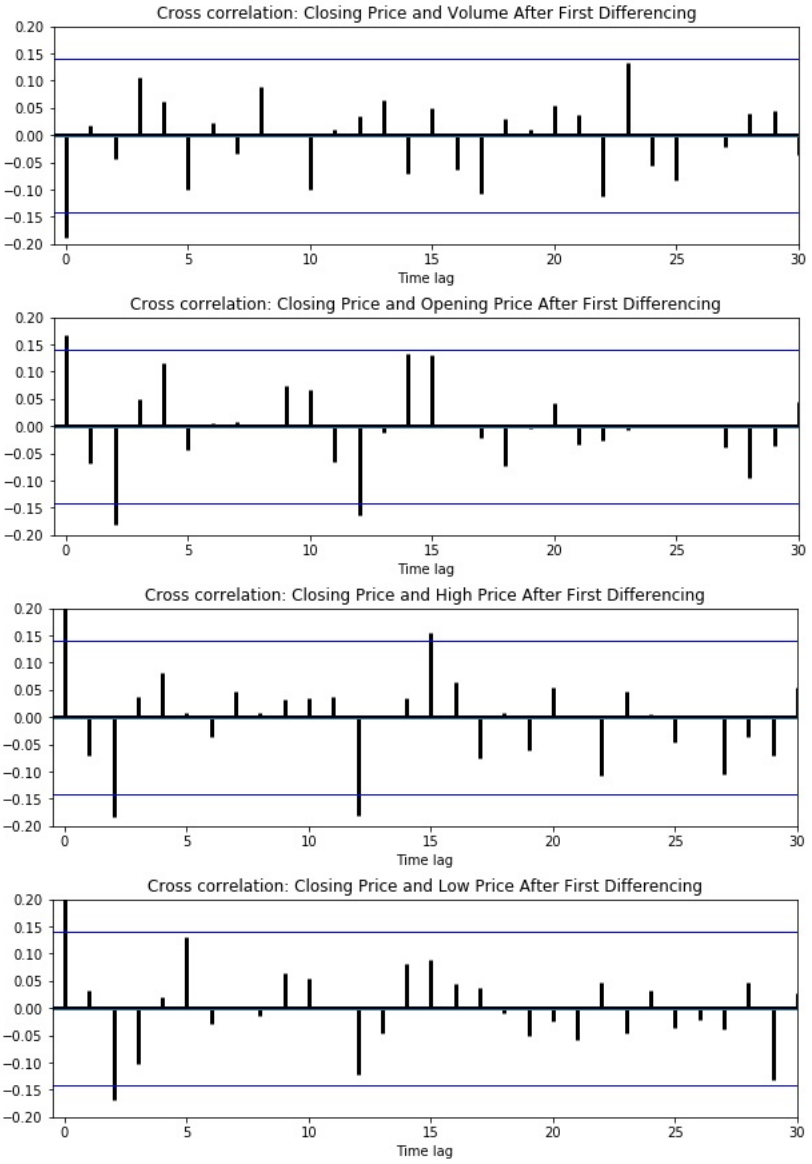

Figure 6: KBANK closing price index

The order (p, d, q) of this model is selected in the same range for ARIMA. We find that the appropriate inputs are opening price and high price which provide the largest cross correlation from two previous day. We put these inputs in the ARIMAX. The data which are use for training and test is the same as ARIMA Model.

ARIMAX	AIC	BIC	Training RMSE
(0, 1, 1)	825	840	3.09
(0, 1, 2)	826	844	3.08
(1, 1, 0)	825	840	3.09
(2, 1, 0)	**826**	**844**	**3.08**
(2, 1, 1)	827	849	3.07
(3, 1, 0)	828	849	3.08
(3, 1, 1)	829	854	3.07
(3, 1, 2)	830	858	3.06

Table 2: Statistical results of ARIMAX models for KBANK stock price prediction

$$A(L)(1-L)y(t) = (1-L)(B_1(L)x_1(t) + B_2(L)x_2(t)) + C(L)e(t) \tag{9}$$

$$A(L) = 1 - A_1L - A_2L^2 - A_3L^3$$
$$B_1(L) = B_{12}L^2$$
$$B_2(L) = B_{22}L^2$$
$$C(L) = 1 + C_1L - C_2L^2$$

where

$y(t)$ is closing price on day t.

$x_1(t)$ is opening price on day t.

$x_2(t)$ is high price on day t.

The result is in Table 2. ARIMAX(2, 1, 0) has low RMSE, AIC and BIC on training set. Therefore, we choose it to be the candidate for model comparison.

3. **ANN Model 1**

First ANN use only closing price on two previous days. It is easy to compare this model to the ARIMA Model, because they use the same inputs. The neural network need enormous data to train. Therefore we use the training set from 1 January 2008 to 28 August 2018 which are 2585 observation. The test set is the same as ARIMA which are 41 observations from 29 August 2018 to 26 October 2018.

$$\hat{y}(t) = f(y(t-1), y(t-2)) \tag{10}$$

where

$\hat{y}(t) \in \mathbb{R}$ is predicted closing price on day t.

y(t) is closing price on day t.

The model architecture which has two input nodes, k node in one hidden layer and one output node is represented as 2-k-1. The number of nodes in hidden layer will be adjusted as shown in Figure 7. The algorithm we use for optimization is ADAM which is stochastic gradient-based optimizer in SciKit Learn. RMSE is not linearly decrease as number of nodes increase, it is volatile because each model converge to different local minimum in each parameters initialization. Then, we compute each RMSE on training set and select the model which has the smallest RMSE. The model which provides the smallest RMSE is 2-42-1.

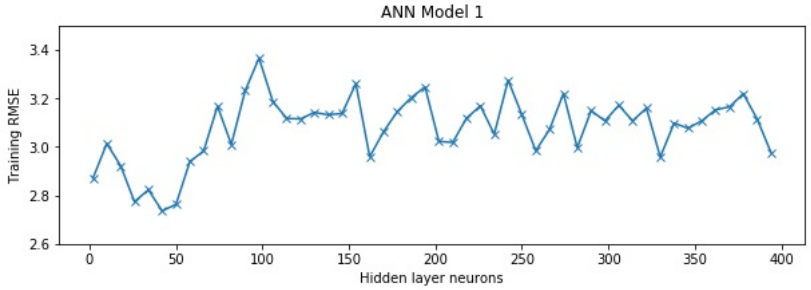

Figure 7: Training RMSE vs Number of nodes in hidden layer

$$RMSE_{train} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{11}$$

where

N is number of samples in training data set.

4. **ANN Model 2**

We use data from two previous days as the inputs of neural networks. Each input vector X(t) contains four values which are opening price, closing price, high price and low price which are the same input as ANN in [1].

$$\hat{y}(t) = f(X(t-1), X(t-2)) \tag{12}$$

where

$\hat{y}(t) \in \mathbb{R}$ is predicted closing price.

$X(t) \in \mathbb{R}^4$ contains opening price, closing price, high price and low price on day t.

f : $\mathbb{R}^8 \to \mathbb{R}$ is nonlinear function of neural networks.

The model architecture which has eight input nodes, k node in one hidden layer and one output node is represented as 8-k-1.The training set and test set are the same as ANN Model 1. The number of nodes in hidden layer will be adjusted as shown in Figure 8. We also use ADAM optimizer in SciKit Learn to estimate this model. The model which provides the smallest RMSE is 8-14-1.

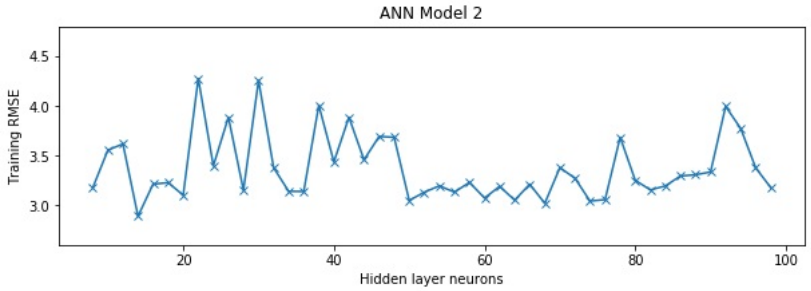

Figure 8: Training RMSE vs Number of nodes in hidden layer

5 Experiments

Training data in time series models and ANN models are different because of different number of parameters in models and training method. We need to compare the models together, therefore we must select the same test sets. In ARIMA and ARIMAX, we split 80% of data to be training set and 20% of data to be test set. The 20% of data are 41 observations from 29 August 2018 to 26 October 2018 and in ANN Model 1 and Model 2 we use 2585 observations from 1 January 2008 to 28 August 2018 for training set and use the same test set as ARIMA.

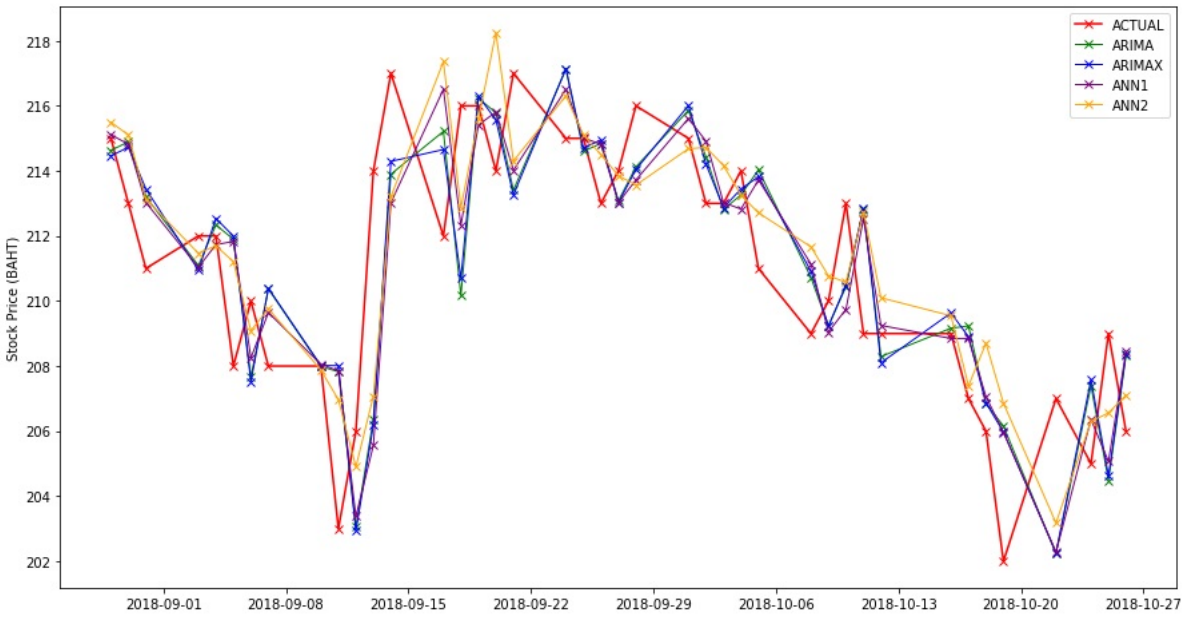

Figure 9: Graphical Representation of actual value and predicted value by each model.

Model	RMSE on test set
ARIMA	2.81
ARIMAX	2.78
ANN Model 1	2.76
ANN Model 2	**2.56**

Table 3: RMSE on test set of each model.

Figure 9 shows that predictions from ARIMA, ARIMAX and two ANNs. In order to evaluate each model precisely, we compute RMSE of all models on test set which are shown in Table 3. ANN Model 2 has the best performance because it is the most complex model we used. ANN Model 1 is second best with the second most complexity. However ANN Model 1 is more complex than ARIMAX, ARIMAX has RMSE which is close to ANN Model 1. Because ARIMAX has more inputs than ANN Model 1. ARIMA Model is the worst compare to all because it has the lowest complexity.

6 Conclusions

This paper show that the superiority of neural networks over time series models. ANN gave the least root-mean-square error on test set because it is the most complex model we used in this research. Neural networks performance can be improved by increasing model complexity. Increasing complexity come with more training data. Because we have a few decades data, the model complexity is limited.

References

[1] Ayodele Ariyo Adebiyi, Aderemi Oluyinka Adewumi, and Charles Korede Ayo. *Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction.* Journal of Applied Mathematics, vol. 2014.

[2] Kaustubh Khare, Omkar Darekar, Prafull Gupta and Dr. V. Z. Attar *Short Term Stock Price Prediction Using Deep Learning.* IEEE International Conference On Recent Trends in Electronics Information and Communication Technology (RTEICT), May 19-20, 2017, India.

[3] Ayodele Ariyo Adebiyi, Aderemi Oluyinka Adewumi, and Charles Korede Ayo. *Stock Price Prediction Using the ARIMA Model.* 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation.

[4] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics).* Springer; 4th ed. 2017 edition (April 11, 2017)